Scientometrics
DOI 10.1007/s11192-012-0930-3

A hybrid keyword and patent class methodology for selecting relevant sets of patents for a technological field

Christopher L. Benson · Christopher L. Magee

Received: 9 August 2012

Abstract This paper presents a relatively simple, objective and repeatable method for selecting sets of patents that are representative of a specific technological domain. The methodology consists of using search terms to locate the most representative international and US patent classes and determines the overlap of those classes to arrive at the final set of patents. Five different technological fields (computed tomography, solar photovoltaics, wind turbines, electric capacitors, electrochemical batteries) are used to test and demonstrate the proposed method. Comparison against traditional keyword searches and individual patent class searches shows that the method presented in this paper can find a set of patents with more relevance and completeness and no more effort than the other two methods. Follow on procedures to potentially improve the relevancy and completeness for specific domains are also defined and demonstrated. The method is compared to an expertly selected set of patents for an economic domain, and is shown to not be a suitable replacement for that particular use case. The paper also considers potential uses for this methodology and the underlying techniques as well as limitations of the methodology.

Keywords Patent searching · Technological planning · Information retrieval · Patent analysis

Mathematical Subject Classification (2000) 68P20

JEL Classification O31 · O32 · O34

C. L. Benson (✉)
Department of Mechanical Engineering, Massachusetts Institute of Technology,
77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: cbenson@mit.edu

C. L. Magee
Engineering Systems Division and Department of Mechanical Engineering,
Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
e-mail: cmagee@mit.edu

Published online: 15 December 2012

Introduction

As technological progress continues to accelerate, a greater need arises to understand how technologies advance over time and how we can implement those lessons into developing technologies for the future. One of the sources of data that has been widely used for understanding technological growth is the patent data that approximately records most of the advances in technology (Campbell 1983). Though the use of patent data for economic and scientific analysis began in 1966 (Trajtenberg 1990), the growing capabilities of computers and data analytics tools have allowed for significant increases in the ability to search the patent data and extract useful information and insights (Joho et al. 2010). The amount of information that is easily accessible through the patent database and a web browser is orders of magnitude higher than what was available just 20 years ago (Michel 2001).

In addition to accessibility, there are other significant reasons why the patent database provides an excellent data source for analyzing technological change over time (Hall and Jaffe 2001). Overall, patents are a set of data that contains the raw information created by the inventors of millions of patents over hundreds of years, and also the underlying information present in the organization of this massive data set that has been created by thousands of expert patent examiners. The combination of the data and organization potentially comprise the 'most valuable data in the world' (Atkinson 2008). More effective use of this powerful patent information for understanding how technology grows over time is enabled if one can develop a robust, repeatable method for finding a relevant and complete set of patents that represent a particular technological field. The relevance of a patent set resulting from a search is defined as the number of relevant patents in that set divided by the total number of patents in the same set. Similarly, completeness is the number of relevant patents in that set divided by the total number of relevant patents in the entire United States patent database. Such a robust, repeatable method is described and evaluated in this paper.

Background

Advances in techniques for patent searching

The most basic ways of searching for patents are the keyword search and the classification search. The keyword search uses search terms and Boolean operators (AND, OR, NOT, NEAR) to construct queries to find the most relevant patents (Larkey 1999). The classification search method requires that the patents already be classified (such as in the US or International Patent classification systems), and that the patent(s) in question can be pinpointed to just one or more patent classes (Baillie 2002). Beyond the two most basic methods for retrieving sets from the patent database, there have been an increasing number of approaches involving complex information retrieval techniques and methods (D'Hondt 2009). Table 1 shows a list of different approaches that have been used by patent researchers in recent years, which altogether makeup a patent searching 'toolbox'.

The techniques in Table 1 are a set of methods that can be combined in different ways to locate a specific set of patents, as demonstrated by Wang (2011). The methodology described and tested in this paper is a novel combination of the two simplest approaches in this listing.

Advances in search methods for technologists and researchers

While there have been many advances in patent searching techniques, there has been very little improvement in the art of broad searches such as ones that would be performed by

Table 1 Patent searching techniques modified from Mahdabi et al. (2011)

Patent searching technique	References
Boolean	Baillie (2002)
US patent classification (UPC)	Baillie (2002)
International patent classification (IPC)	Takaki et al. (2004)
Query expansion	Wang (2011)
Dividing into different time periods	Wang (2011)
Probabilistic retrieval models	Fujita (2004)
Citation linking	Fujii (2007); Lopez and Romary (2010)
Unigram and bigram frequency analysis	Magdy and Jones (2010)
Knowledge representations of data	Graf et al. (2010)
Using sample patent to generate keywords	Xue and Croft (2009)
Semantic analysis	Gerken and Moehrle (2012)

academics, economists, or those looking for a general overview of a technological field. Atkinson (2008) discusses how little the methods for searching the patent database have changed in recent times, stating that

> *'The Basic Way to search (mostly Boolean in logic structure, even if natural language has been used as a nexus) has changed little since … 25 years ago)… We have far more databases available, more beautiful and comprehensive results display, but getting those hits still relies on set theory and exclusion.'*

An important case in the study of technological development is the work done by Trajtenberg (1987) in his analysis of computed tomography (CT) patents. Trajtenberg describes the use case of his set of 456 CT patents and importantly establishes that higher cited patents have higher value. To establish his patent set, he used his extensive case study of the CT industry (companies, installations in hospitals, inventors, etc.) to supplement a word search to find patents. He also read all the abstracts in his patent set to exclude inappropriate patents. Trajtenberg describes his method as one of trial and error in which he uses a variety of different aspects of the patent including the classification codes, assignees and regular Boolean search:

> *'The search for patents in a particular product field or industry can be done in a variety of ways: using key words pertaining to the product in question that may appear in the title and/or in the abstract, identifying a small set of relevant patent classification codes, locating assignees (typically firms) that are known to operate in the field,* etc. *Needless to say, there isn't a well-defined method that would deliver with certainty* ***all*** *the patents in a given field, and* ***only*** *those.' (Emphasis his)* (Trajtenberg 1987)

While Trajtenberg's method resulted in a patent set that is certainly more appropriate for his purposes than any others yet demonstrated, it is not clear that it can be reproduced in other technological domains and in fact the approach has not yet been applied anywhere else. It requires extensive knowledge of terminology in the field as well as information about relevant firms and he even detailed every installation of CT during his search, which was limited to 1971–1986. Our aim is to create a repeatable and simpler to use method in order that a user of the hybrid keyword-classification (HKC) method can quickly and easily

compare patent sets across many technical fields of interest over a longer period of time than Trajtenberg considered. We do not anticipate any such simple method to reproduce what Trajtenberg retrieved as much context he used is lost; however, a supplemental procedure should have value nonetheless.

The direction we have chosen to follow in developing a supplemental approach was pointed to by Atkinson (2008) when she discussed the "need for growth away from reliance on words and language and to draw in tools having defined quantitative values such as patent classes."

The hybrid keyword-classification patent searching method

In order to address the aforementioned opportunities, we sought to create a method that was easily repeatable so that it can be used quickly by many different types of users, including those who are not well versed in the complexity of the patent classification systems. The method was designed to give a relatively complete set of patents that are relevant to a particular technological domain. The goal was to create a data set that had high completeness, meaning that it included a high proportion of the total number of relevant patents. The data set must also have high relevancy, so that it includes relatively few non-relevant patents. Finally, the methodology should be robust and flexible to meet the varying needs of different users. This section will describe the intuition behind the new HKC patent searching method and will describe in detail how the method works. The key concept is the addition of a new cross-classification tool, so the paper considers its strengths and weaknesses, and how it can be considered another technique in the 'toolbox' of patent searching.

Pre-search US issued patent titles and abstracts for the search term

The first step of the HKC is to pre-search using a set of keywords to begin the process of finding the most representative patent classes (in both the US and International Patent systems), which is defined in the following section.

As one of the goals of the method was to be simple and easy to use even for someone not fluent in the patent system, the input to the HKC is simply a set of search terms that can be entered into a text box. This works best with search queries of two words (ex: solar photovoltaic), which suits our use case of technological development research. The pre-search was completed by searching for the two-word query in the title or the abstract of United States Issued Patents. Thus, the pre-search identifies a set of patents with the specific query in the title or abstract.

The pre-search was done using the patent search tool PatSnap (2012), which searched all US patents from 1971 to the present and was used as our database for further analysis (all of the searches in this paper were completed in May of 2012 unless noted otherwise). In this paper we will give the search queries that can be used in http://www.patsnap.com/patents because it is publicly available and has a faster startup time than recreating a patent database from scratch. The search query used for the pre-search for 'solar photovoltaic' at http://www.patsnap.com/patents is:

'TTL: (solar photovoltaic) OR ABST: (solar photovoltaic) AND DOCUMENT_TYPE: United States Issued Patent'

This search returns 991 patents.

Rank the IPC and UPC patent classes that are most representative of the technology

The next step in the HKC method is to use the set of patents resulting from the pre-search to determine the US patent classes (UPC) and international patent classes (IPC) that are most representative of the specific technology. The representativeness ranking for the patent classes is accomplished by using the mean-precision-recall (MPR) value. This value was inspired by the 'F1' score that is common in information retrieval, but uses the arithmetic mean (instead of the geometric mean) of the precision and recall of a returned data set (Magdy and Jones 2010). Table 2 shows an example MPR calculation for the UPCs and IPCs in the pre-search for 'solar photovoltaic'. In the paragraphs below we will describe the calculations to arrive at each column in this table and will use UPC 136 as the example.

Using the set of patents from the pre-search, we determine all of the unique patents classes that appear in the set. For example, within the pre-search results for 'solar photovoltaic', there are 22 unique IPCs and 10 unique UPCs. Table 2 lists the five IPCs and UPCs (column 1) with the most patents present in the search for 'solar photovoltaic'. The number of patents identified in the pre-search that are present is each class is shown in column two (this can also be called the overlap of the pre-search and patent class); it is found using a search similar to the following (using UPC 136 in this example, which returns 608 patents):

'CCL: (136) AND TTL: (solar photovoltaic) OR ABST: (solar photovoltaic) AND DOCUMENT_TYPE: United States Issued Patent'

Note that the sum of column two is often greater than the total number of patents in the pre-search group due to the fact that many patents are classified in multiple UPCs or IPCs.

Next, we are interested in computing the fraction of the patents in the pre-search that fall within each patent classification, also called the patent class Recall and shown in column 3 of Table 2. The recall for each of the listed patent classes is calculated by dividing the number of patents in the pre-search results that are within the patent class (column 2) by the number of patents in the pre-search patent set (991 for the example of 'solar photovoltaic'). For UPC 136, the recall is $608/991 = 0.61$.

Table 2 Example calculation of MPR for five UPCs and five IPCs for the search term 'solar photovoltaic'

Patent class	Number of patents in pre-search and patent class	Patent class recall (column 2/991)	Total Number of patents in patent class	Patent class precision (column 2/column 4)	MPR (column 3 +column 5)/2
UPC-136	**608**	**0.61**	**7,489**	**0.081**	**0.34**
UPC-257	342	0.35	170,333	0.002	0.17
UPC-438	178	0.18	108,686	0.002	0.09
UPC-126	131	0.13	15,450	0.008	0.07
UPC-52	66	0.07	182,645	0.000	0.03
IPC-H01L	549	0.55	180,204	0.003	0.28
IPC-F24J	78	0.08	3,751	0.021	0.05
IPC-H02N	47	0.05	2,133	0.022	0.03
IPC-H02J	40	0.04	10,876	0.004	0.02
IPC-E04D	27	0.03	3,209	0.008	0.02

The calculation for UPC 136 is described in the text and is bold in this table. Note that the pre-search returned 991 patents

$$Recall = \frac{\#\ Patents\ in\ the\ Presearch\ within\ the\ Patent\ Class}{\#Retrieved\ Patents\ in\ the\ Pre-Search}$$

Next, we want to determine the total size of each of the patent classes of interest. Column number 4 shows the total number of patents in each patent class, which is found by the following search (using UPC 136 as the example, which returns 7489 patents):

'CCL: (136) AND DOCUMENT_TYPE: United States Issued Patent'

Given the total size of the patent class, we determine the fraction of the patents in each patent class present in the pre-search, which is called the patent class precision (column 5). This normalizes the weight of very large and very small patent classes that may be over or under represented in the pre-search due to their different sizes. Calculate the precision of each patent class within the pre-search by dividing the number of patents in both the search and the patent class (column 2) by the total number of patents in the patent class (column 4). For UPC 136, the precision is 608/7489 = 0.081.

$$Precision\ =\ \frac{\#\ Patents\ in\ the\ Presearch\ within\ the\ Patent\ Class}{\#\ Patents\ in\ patent\ class}$$

Finally, we find the mean of the precision and recall values, which gives us an estimate of how well each patent class represents the pre-search set. The MPR of each patent class (column 6) is calculated by taking the mean of the patent class precision (column 5) and patent class recall (column 3). The MPR for UPC 136 is (0.68 + 0.081)/2 = 0.34.

$$MPR\ =\ (Precision\ +\ Recall)/2$$

The MPR for each potentially representative patent class—identified by containing patents present in the pre-search—are then ordered from highest MPR to lowest for both the IPC and UPC patent classification systems.

Select the overlap of the most representative IPC class and UPC class

To find the final set, the patents that are contained within both the IPC and UPC classes with the highest MPRs within the set of US issued patents are retrieved. Our intuition for this step is founded upon the extensive patent examiner experience and knowledge embedded in these two classification systems: the concept is to utilize that embedded knowledge. If a patent is listed in the most representative patent class in both systems (particularly since the two systems are somewhat differently structured), a reasonable hypothesis is that such dual membership results in obtaining patents of higher relevance (Criscuolo 2006). With patents having multiple entry systems, the completeness of the set may at the same time not be too compromised. The results section will test this intuition but first we complete our description of the method.

For the solar photovoltaic case, Table 3 shows the top two classes for the IPC and UPC with their corresponding MPRs as well as the size of the returned data set when the overlap of the two classes was retrieved. For example, the number of patents simultaneously contained within both highest ranked classes [136 (UPC) and H01L (IPC)] is 5101, whereas the overlap of 257 and F24J is only 16 patents, indicating quite low completeness. Selecting classes with high MPRs generally results in higher completeness and relevancy percentage combinations in the final patent set. For example, the large set ($n = 136406$) obtained when patents that are contained within both H01L and 257 consists of a very large fraction ($\sim$98 %) of irrelevant patents.

Table 3 Comparison of top IPC and UPC classes for the search term 'solar photovoltaic'

'Solar photovoltaic'	H01L (MPR = 0.28)	F24 J (MPR = 0.05)
136 (MPR = 0.34)	5,101	260
257 (MPR = 0.17)	136,406	16

The 4 entries in the lower right hand boxes is the number of patents that are simultaneously listed in both of the specified classes (the overlap of the two classes)

We will later discuss modifications but our direct method is to select the patents that are in both the most representative UPC and in the most representative IPC. For the solar photovoltaic example, the patent set obtained from the overlap of the most representative classes (136 and H01L) is obtained by the following query at http://www.patsnap.com/patents.

'CCL: (136) OR ICL: (H01L) AND DOCUMENT_TYPE: United States Issued Patent'

Test the resulting patent set for relevancy

Although in this demonstration case we tested some preliminary sets for relevance, the basic process involves performing the relevance test (done by reading the abstracts of a random or semi-random test set of patents) after obtaining the set from crossing the most representative UPC class with the most representative IPC class. The relevancy sample test set size for all larger sets of patents should be 300 patents to ensure a 95 % confidence interval with a margin of error of 5 %. The test set structure can be varied for different purposes. For example in the case of study of technological change in a domain, we are very interested in the most highly cited patents, therefore we took the top 100 most highly cited patents and added another 200 randomly selected patents for the test set and

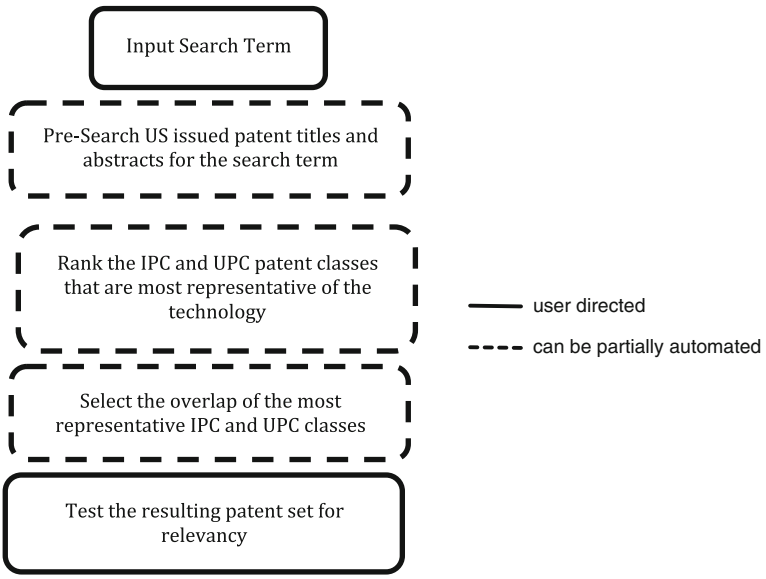

Fig. 1 Overview of the HKC method: most of the method can be automated via a computer, with only the selection of the search query and the testing of the final results left to the user

determined the number of relevant patents using a methodology discussed in a previous paper (Benson and Magee 2012).

Depending on the use case, there may be a need for a various levels of relevance up to 100 % but we consider greater than 60 % of the patents being relevant acceptable for broad study of technological change. It is also important to note the absolute size of the patent sets; we found that this number varied considerably. In studying technological change over time in a domain, we are more tolerant of non-relevant patents as long as we retrieve >75 % of the relevant patents (high completeness is favored over high relevance), but that may not be the case for all uses.

Figure 1 shows a summary of the HKC method and the use of the keyword and classification system to gain more complete and relevant patent sets.

Results

The HKC is a method designed for ease of use and to retrieve relatively complete and relevant sets of patents. In order to explore the effectiveness of the method, we have tested it against other methods. The following sections show the results of several comparative tests for the effectiveness of the HKC method.

Overview of cases

The most important measure of effectiveness for the HKC method is if it can provide highly relevant and complete patent sets for the user. Table 4 compares the overall size (an indicator of completeness) and relevancy percentage (based upon the sampling method described above) of the returned search results of three different methods across five technological domains. We compare the direct HKC results with those resulting from a keyword search in the title and abstract because searching the total patent leads to very low relevancy and searching the title alone gives very poor completeness. Specifically for the comparative keyword searches in Table 3 we use:

'ttl: (keyword) OR abst: (keyword) AND (DOCUMENT_TYPE: United States Issued Patent)'

This query in fact is equivalent to our "pre-search" but for the keyword search method, the resulting set of patents is the "final" set (we will examine the impact of differing search terms shortly as this has significant impact on the keyword search patent set). Table 3 also compares the patent set that is achieved by the method of UPC classification selection. Typically, the patent classification method involves examining the UPC classification titles and making a subjective judgment on which class is best for the field of interest or possibly subjectively defining several classes of potential interest. In our comparison in Table 3, we wanted a stable method so the UPC classification selected was the most representative UPC class for each field based on the objective MPR method defined above. For example, UPC class 136- 'Batteries: thermoelectric and photoelectric' was the class identified for the 'photovoltaic electricity' case.

The results show that the HKC method does not always simultaneously produce the highest relevancy percentage and the highest estimated completeness (relevancy times size), but it consistently performs well for both characteristics and does not yield very poor results as often occurs when using the keyword or classification selection methods.

The HKC search for 'photovoltaic electricity' provides a significant improvement over the authors' last attempt, which resulted in a set of 2484 patents with only 62 % relevancy

Table 4 Size and relevancy (in brackets) of returned data sets for three search methods

Field of interest	HKC	Keyword	Classification selection (UPC)
Photovoltaic electricity	5,101 (85 %)	1,006 (75 %)	7,233 (57 %)
Wind turbine	1,346 (94 %)	1,843 (91 %)	12,893 (26 %)
Electric capacitor	6,173 (84 %)	11,026 (43 %)	9,472 (2 %)
Electrochemical battery	22,115 (62 %)	1,159 (87 %)	26,111 (62 %)
Computed tomography	3,827 (91 %)	1,289 (98 %)	10,444 (69 %)

using a non-automated but more elaborate keyword search technique (Benson and Magee 2012). In Table 4, it is noteworthy that the HKC method starting from the same keyword search terms results in a patent set that is five times larger and of higher relevancy than the keyword search. The HKC photovoltaic set of patents is also superior to the patent set from the third method in Table 3 as seen by relatively lower relevance for the classification selection set. The results of the keyword search for 'Wind Turbine' are marginally superior to those of the HKC method, but the keyword search produces *much* less complete patent sets for three of the other fields of interest. Similarly, the classification selection method produces somewhat superior results for the 'electrochemical battery' query but remarkably poor relevance for capacitors. In the CT set, the HKC produces a moderately large and very relevant set of patents but does not appear to be as complete as the classification selection method.

Flexibility of search terms and robustness of HKC method

The HKC method requires only a two-word search term that describes the technology of interest and can take multiple synonymous or near-synonymous queries and will give the same result. For example, the search queries 'solar power' and 'photovoltaic electricity' provide the same end patent sets with HKC method but very different results if one just uses a keyword search. Table 5 shows a comparison of the robustness of the HKC and keyword search methods across different search terms.

Table 5 indicates that the HKC method has a low sensitivity to the selection of initial search terms as only the term dielectric capacitor led to a substantially different set. On the other hand, the same keyword differences lead to substantial differences in the keyword search method for almost all cases. Thus, to learn more about solar photovoltaic technology, HKC offers a stable (and relatively complete and relevant) patent set using a variety of different search terms whereas the keyword search data sets would be variable and of unknown quality. This lack of sensitivity to specific search terms indicates that the HKC is more repeatable across different users and technical domains.

Modifications to the HKC method

In arriving at the results in Table 4, the authors noted some further useful information that suggests modifications of the method for specific fields. For the patent set in the 'wind turbine' query, the 416 and 290 UPCs are almost equally representative (MPR for $416 = 0.45$, MPR for $290 = 0.36$). Such close comparisons occur rather often, but it many cases most of the similarly representative patent class is almost entirely present in the directly determined patent set. For example, 88 % of the patents in the 204/H01M overlap

Table 5 Comparison of results of various search queries using the HKC and keyword searches

Search term	HKC (UPC/IPC)	HKC # patents	Keyword (# patents)
Photovoltaic	136/H01L	5,101	3,000
Solar power	136/H01L	5,101	2,443
Solar photovoltaic	136/H01L	5,101	896
Photovoltaic electricity	136/H01L	5,101	1,014
Wind turbine	416/F03D	1,367	1,875
Windmill	416/F03D	1,367	414
Aerogenerator	290/F03D	1,103	10
Electrochemical battery	429/H01M	22,115	1,159
Electrochemical cell	429/H01M	22,115	5,352
Secondary battery	429/H01M	22,115	3,654
Electrical capacitor	361/H01G	6,173	11,049
Dielectric capacitor	257/H01G	850	7,845
Computed tomography	378/A61B	3,827	1,288
Cat scan	378/A61B	3,827	62

are present in the 429/H01M overlap in the search for 'electrochemical cell' (MPR for 204 = 0.14, MPR for 429 = 0.37, MPR for H01M = 0.41). In the case of the wind energy example, there is only a 30 % redundancy between the 416/F03D overlap and the 290/F03D overlap, but both patent sets are relevant to the wind energy generation field. Specifically, the 416/F03D overlap has patents that are related to the blades of a wind turbine, while the 290/F03D overlap contains patents primarily involved in the gearbox and generator portion of the wind turbines. This is also indicated by the UPC titles for the patent classes: 416—'Fluid reaction surfaces (i.e., impellers)' and 290—'Prime mover dynamo plants'. In this case, for further analysis of technological change, we recommend using both the 290/F03D overlap and the non-redundant part of the 416/F03D overlap, which, when combined, result in one patent set containing 2,078 patents.

This same technique was used for obtaining a CT patent set where the 378/G01N overlap, which includes 3,814 patents with 76 % relevancy, is combined with the original set of 378/A61B (3,827/91 %) to create a final data set with 7,330 patents with 84 % relevancy The appropriate query for the CT search is:

"CCL: (378) AND (ICL: (A61B) OR ICL: (G01N)) AND (DOCUMENT_TYPE: United States Issued Patent)"

Another emendation suggested from the relevance test experience is further pruning of a particular overlap after the fact. This is demonstrated clearly by the search for patents related to energy storage batteries, which results in the cross of 429—'Chemistry: electrical current producing apparatus, product, and process' and H01M—'Processes or means, e.g., batteries, for the direct conversion of chemical energy into electrical energy'. The unaltered set—see Table 3—results in 62 % relevancy, which is only marginally adequate for our use case. During the relevancy sampling, it became clear that many of the non-relevant patents were related to fuel cells. Therefore, in order improve the patent set, we simply removed many fuel cell patents from the set by eliminating patents with fuel cell in the title -using the following query on http://www.patsnap.com/patents:

"((CCL: (429) AND ICL: (H01M)) NOT (TTL: (Fuel Cell))) AND (DOCUMENT_TYPE: United States Issued Patent)"

Table 6 Comparison of the different search methods including the adjusted HKC method

Field of interest	HKC	HKC modified	Keyword	Classification selection
Photovoltaic electricity	5,101 (85 %)	5,101 (85 %)	1,006 (75 %)	7,233 (57 %)
Wind turbine	1,346 (94 %)	2,078 (94 %)	1,843 (91 %)	12,893 (26 %)
Electric capacitor	6,173 (84 %)	6,173 (84 %)	11,026 (43 %)	9,472 (2 %)
Electrochemical battery	22,115 (62 %)	16,466 (83 %)	1,157 (87 %)	26,111 (62 %)
Computed tomography	3,827 (91 %)	7,330 (84 %)	1,289 (98 %)	10,444 (69 %)

The removal of the patents that had fuel cell in the title resulted in a reduction of 5,649 patents, leaving a data set of 16,466 patents, which, when sampled had a greatly improved 83 % relevancy. This emendation helps alleviate any issues arising from very large patent classes in either the IPC or UPC with relatively small amounts of extra work by the user. Table 6 shows the comparison of the effectiveness of the different search methods including the HKC method with modifications for the three fields just discussed.

When including the emendations beyond the fully automated HKC method, the method becomes better than the other search methods across all of the queries we tested. While the modifications to the original HKC method are helpful, they are certainly not necessary if one is interested in searching for a large number of data sets across many technical fields, as can be the case for some research related to technological development. However, the ease of making such modifications does enhance the usefulness of the HKC method.

Comparison to an expert selection of patents

While the HKC method is not intended to be a replacement for an expertly selected set of patents, it is useful to understand how the method compares with a set of patents hand selected by an expert. In order to do this, the patents from the three CT searches are compared with Trajtenberg's set, using 1973–1987 as the search years in order to match the years spanning Trajtenberg's search. The results are shown in Table 7.

The HKC set contains 524 patents, with 136 patents overlapping Trajtenberg's set of 456 patents. If we were to assume that Trajtenberg's set is the complete set of relevant patents for the CT economic domain, the HKC method would have a relevancy of 26 % and a completeness of 30 %. The keyword search is far worse in completeness and the classification method is far inferior in relevance so the HKC is the best of three weak comparators in this case. While the HKC method does not match up well with the Trajtenberg set, it does manage to locate four of the top five most cited patents in the Trajtenberg set as well three other highly cited patents that were not in Trajtenberg's set. The results look more promising when the adjusted HKC method including the 378 and G01N data set is used, thereby locating seven of the top ten most cited patents in Trajtenberg's set.

The highly cited patents in the HKC results that were not found in Trajtenberg's data set highlight the difference between searches within a technical or an economic category. Trajtenberg aimed to:

> *'allow one to identify quite easily all the patents issued in predetermined economic categories, and retrieve them for further analysis.'* (Trajtenberg 1987)

Table 7 Comparison of search methods to Trajtenberg's expert set (limited to years 1973–1987)

	HKC	HKC modified	Keyword	Classification selection
Number of patents	524	1,373	113	3,812
Overlap with Trajtenberg	136	239	76	426

While his analysis is very similar to that of ours, we are primarily interested in identifying patents in a predetermined *technical* category as opposed to the *economic* domains that Trajtenberg focused upon. For example, the highly cited patent-number 4583242 'Apparatus for positioning a sample in a computerized axial tomographic scanner'—describes a method for locating a core sample from a borehole. This patent uses CT outside of medical applications for identifying samples of rock or core for the petroleum industry. However, the patent still represents a development within the technological field. Patents such as these are clearly outside of Trajtenberg's intended field of study, but are within our broader field of study, as we are concerned with including technological spillover in our studies (Benson and Magee 2012).

Ultimately the marginal agreement of the patents found by the HKC method and the Trajtenberg set demonstrates that the HKC method alone is not a replacement for an expertly selected set of patents within an economic category, but rather a robust tool to be used in conjunction with others to locate a set of patents relevant to a particular technical field.

Discussion

This paper introduced the problem of locating relatively complete sets of patents relevant to a specific technical field in a simple and repeatable manner. The hybrid keyword-classification method provides a new methodology to locate highly relevant and complete sets of patents within a technological field. The method is easily automated and straightforward to use, only requiring a query related to the field of interest. Moreover, our results show that the HKC method allows for flexibility in the initial keywords chosen. We have shown that the patent sets obtained from the HKC method are nearly always an improvement over those obtained from the keyword or the classification search methods. Importantly, the HKC method is more robust and generally easier to use. The method acts as a supplement, not a replacement for an expertly selected set of patents. HKC is a simple and repeatable method for selecting sets of patents relevant to a particular technical field. The repeatability of this method should help improve consistency of patent analysis across many fields.

Limitations and future research

While the method is almost always better than the equally simple alternatives, there are some cases where the HKC method does not provide a result better than the keyword or classification search. We have introduced a set of still simple modifications that make the HKC method more effective and the results from this modified HKC method appears generally superior to the other simple approaches. Whether or not to use the emendation will depend on the use case for the resulting data set as they require a tradeoff of automation for effectiveness.

Another limitation of the HKC method is the level of technical hierarchy that can be used effectively in the initial search query. Search queries that are more generic (i.e., electricity) or are more specific (i.e., mono-crystalline silicon photovoltaic electricity) do not produce effective results in this incarnation of the HKC method. We believe it is possible to alter the specificity in search terms and determine a set of more specific classifications (H vs. H01L vs. H01L 21/02) that represent a technology and this is something that is recommended for future research. As an example, a search for 'luminescent solar cells' indicates the overlap of UPC 136/247 and IPC H01L31/04; which produces a set of 35 patents, 34 of which were related to luminescent solar cells.

The most immediate future use development of the HKC will be in further use in the research of technological development such as in Benson and Magee (2012). Nevertheless, the core idea of finding relevant information by a combination of a keyword and a classification system is one that can be applied in areas other than just the intellectual property domain. There is the possibility to apply the general approach to areas such as science, healthcare and law where multiple expert maintained classification systems seem to exist.

Acknowledgments We are pleased that Professor Trajtenberg still had access to his patent set and grateful that he cooperatively shared it with us. We also thank Subarna Basnet for useful input on an earlier draft. The research was supported by the SUTD/MIT International Design Center.

References

Atkinson, K. (2008). Toward a more rational patent search paradigm. *Proceedings of the 1st ACM workshop on Patent Information Retrieval*, pp. 37–40.

Baillie, J. (2002). Introduction to patent searching. Resource document. Boston Public Library. http://www.bpl.org/research/govdocs/patent_handout.pdf. Accessed 25 May 2012.

Benson, C. L., & Magee, C. L. (2012). A framework for analyzing the underlying inventions that drive technical improvements in a specific technological field. *Engineering Management Research, 1*(1), 2–14. doi:10.5539/emr.v1n1p2.

Campbell, R. (1983). Patent trends as a technological forecasting tool. *World Patent Information* (1979), 137–143.

Criscuolo, P. (2006). The 'home advantage' effect and patent families. A comparison of OECD triadic patents, the USPTO and the EPO. *Scientometrics, 66*(1), 23–41.

D'hondt, E. (2009). Lexical issues of a syntactic approach to interactive patent retrieval. *The Proceedings of the 3rd BCSIRSG Symposium on Future Directions in Information Access*. pp. 102–109.

Fujii, A. (2007). Enhancing patent retrieval by citation analysis. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. doi: 10.1145/1277741.1277912.

Fujita, S. (2004). Revisiting the document length hypothesis – NTCIR-4 CLIR and patent experiments at Patolis. *Proceedings of NTCIR-4 Workshop*.

Gerken, J. M., & Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics, 91*(3), 645–670. doi:10.1007/s11192-012-0635-7.

Graf, E., Frommholz, I., Lalmas, M., & van Rijsbergen, K. (2010). Knowledge modeling in prior art search. *Advances in Multidisciplinary Retrieval*. pp. 60–71.

Hall, B., & Jaffe, A. (2001). The NBER patent citation data file: lessons, insights and methodological tools. *NBER Working Paper Series*, p. 8498.

Joho, H., Azzopardi, L., & Vanderbauwhede, W. (2010). A survey of patent users. *Proceedings of the third symposium on Information interaction in context*. pp. 13–22.

Larkey, L. (1999). A patent search and classification system. *Proceedings of the fourth ACM conference on Digital Libraries*. pp.179–187. doi:10.1145/313238.313304.

Lopez, P., & Romary, L. (2010). Experiments with citation mining and key-term extraction for prior art search. *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*.

Magdy, W., & Jones, G. J. F. (2010). PRES: a score metric for evaluating recall-oriented information retrieval applications. *SIGIR '10 Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. pp. 611–618.

Mahdabi, P., Keikha, M., Gerani, S., & Landoni, M. (2011). Multidisciplinary Information Retrieval. Building queries for prior-art search (pp. 3–15). Berlin, Springer.

Michel, J. (2001). Patent citation analysis. A closer look at the basic input data from patent search reports. *Scientometrics, 51*(1), 185–201.

Patsnap. (2012). Patsnap patent search and analysis. Retrieved May 15, 2012, from http://www.patsnap.com.

Takaki, T., Fujii, A., & Ishikawa, T. (2004). Associative document retrieval by query subtopic analysis and its application to invalidity patent search. *Proceedings of the 13th ACM conference on Information and knowledge management-CIKM'04*, p. 399.

Trajtenberg, M. (1987). Patents, citations and innovations: tracing the links. *NBER Working Paper Series*, p. 2457.

Trajtenberg, Manuel. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics, 21*(1), 172–187.

Wang, S.-J. (2011). The state of art patent search with an example of human vaccines. *Human Vaccines, 7*(2), 265–268. doi:10.4161/hv.7.2.14004.

Xue, X., & Croft, W. B. (2009). Automatic query generation for patent search. *Proceeding of the 18th ACM conference on Information and knowledge management-CIKM'09*, p. 2037.